E-033

日本語における大規模言語モデルを用いた共感応答生成での常識知識グラフ活用の検討

A study on Incorporating Commonsense Knowledge Graphs into Japanese Empathetic Response Generation Using Large Language Models

久保 晴生 †, 田村 晃裕 †, 加藤 恒夫 †
Haruki Kubo, Akihiro Tamura, Tsuneo Kato

1 はじめに

自然言語処理の分野では，ユーザに寄り添った対話システムを実現するため，ユーザの感情に共感を示す応答を自動生成する共感応答生成の研究が古くから盛んに行われており，近年では，大規模言語モデル (LLM) を用いた手法が主流となっている．英語における共感応答生成においては，イベントに対する人間の日常的な常識や感情の因果関係を表現する常識知識グラフを活用することで，応答の質が向上することが示されている．例えば，Sabour ら [2] は，Transformer エンコーダ・デコーダモデルに基づく共感応答生成において，対話履歴のエンコード時に常識知識グラフ ATOMIC2020 を活用することで応答の共感性を改善した．また，Qian ら [4] は，LLM に基づく共感応答生成において ATOMIC2020 の有効性を示した．しかし，日本語における共感応答生成では常識知識グラフを活用する試みは行われておらず，その有用性は確認されていない．

そこで本研究では，日本語において，常識知識グラフを活用した LLM による共感応答生成モデルを構築し，日本語の共感応答生成における常識知識グラフの有用性を検討する．具体的には，日本語の常識知識グラフ Atomic-ja[5] に基づき，直前のユーザ発話に関する 4 種類の関係（必要，影響，意図，反応）の推論を生成し，生成した推論を含めたプロンプトを用いて LLM により共感応答を生成するモデルの性能を評価する．また，Qian ら [4] により提案された 2 段階手法と Few-shot 手法を日本語の共感応答生成で実装し，これら手法と常識知識グラフを活用した手法を組み合わせることの有効性も検証する．

2 提案モデル

本節では，常識知識グラフを活用した LLM による日本語の共感応答生成モデルを提案する．提案モデルの概要を図 1 に示し，用いたプロンプトを図 2 に示す．なお，図 2 中の {context} には応答対象の対話履歴が入る．

2.1 常識知識グラフを活用したモデル（知識ベース）

知識ベースモデルは，Atomic-ja に基づき，対話履歴の最後の発話に関する 4 種類の関係（必要，影響，意図，反応）の推論を生成し，生成した推論を含めたプロンプトを用いて LLM により共感応答を生成する．ここで，「必要」は前にその人に起こっていただろうこと，「影響」は後にその人に起こるだろうこと，「意図」は前にその人が思っていただろうこと，「反応」は後にその人が思うだろうことである．そして，Atomic-ja はイベントの常識知識を「イベント，関係，推論」の三つ組で表現するタプル集合である．

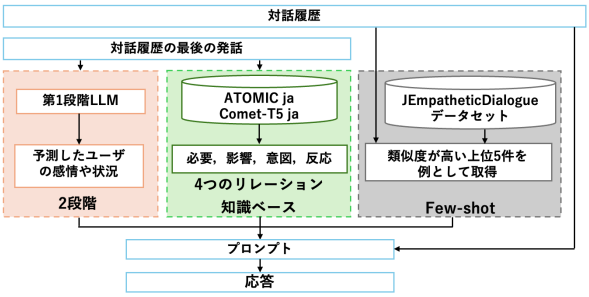

図 1: 提案モデルの概要図

{'role': 'system', 'content': ' これは共感的な対話タスクです。最初の作業者（話し手）は感情のラベルを与えられ、その感情を感じた状況について自身の説明を書きます。その後、話し手はその状況をもう一人の作業者（聞き手）との会話の中で語ります。話し手の感情ラベルや状況は、聞き手には見えません。聞き手は、対話の中で他者の感情を認識し、できる限り共感を示す必要があります。あなたはこれから聞き手の役割を担い、与えられた文脈に基づいて適切な応答をしてください。聞き手としての次の発言のみを提供してください。
以下は常識推論を活用し、推論されたユーザーの発言や状況に関連する背景情報です。
{knowledge}
以下はユーザーの発言をもとに、ユーザーの感情もしくは状況を推測したものです。
{twostage}
以下は類似の状況での対話例です。
{fewshot}
対話履歴
{context}
これらの情報をもとにあなたは聞き手の役割を担い、既存の文脈に応じた適切な返答をしてください。次の聞き手の応答のみを提供してください。}

図 2: プロンプト（知識 +2 段階 +Few-shot）

知識ベースモデルは，まず，日本語版 T5 モデルを Atomic-ja でファインチューニングした常識推論モデル COMET-T5 ja[1]を用いて，対話履歴の最後のユーザ発話に関する 4 種類の各関係の推論を生成する．そして，生成した推論をプロンプト（図 2 の {knowledge} 部分）に追加し，そのプロンプトを用いて gpt-4o により応答を生成する．常識知識グラフに基づいて生成した推論を用いること

† 同志社大学大学院 理工学研究科　Graduate School of Science and Engineering, Doshisha University

[1] https://huggingface.co/nlp-waseda/comet-t5-base-japanese

表 1: 各共感応答生成モデルの性能

	共感性	一貫性	情報性	流暢性
ベースライン	4.54*	5.00	2.90	5.00
知識	**4.78**	5.00	2.74	5.00
2 段階	4.72	5.00	2.96	5.00
Few-shot	4.44*	5.00	2.38*	5.00
知識 +2 段階	4.70	5.00	**3.00**	5.00
知識＋ Few-shot	4.48*	5.00	2.72	5.00
2 段階＋ Few-shot	4.50*	5.00	2.80	5.00
知識 +2 段階 ＋ Few-shot	4.58*	5.00	2.88	5.00

で，応答生成時にユーザの感情やその原因の理解が促進され，共感性の改善が期待できる．

2.2 2 段階手法や Few-shot 手法との統合モデル

本節では，Qian ら [4] の 2 段階手法や Few-shot 手法を日本語に適応したモデルを 2.1 節の知識ベースモデルと組み合わせたモデルを提案する．2 段階手法では LLM（gpt-4o）を 2 回使用する．まず，1 段階目の LLM で，対話履歴の最後のユーザ発話を基にユーザの感情や状況を予測する．その後，1 段階目で予測されたユーザの感情や状況をプロンプト（図 2 の {twostage} 部分）に追加し，そのプロンプトを用いて 2 段階目の LLM で応答を生成する．

Few-shot 手法では，日本語の共感的対話データベース JEmpatheticDialogue[1] から抽出した共感応答の例を使用する．具体的には，まず，JEmpatheticDialogue の訓練データから応答対象の対話履歴と類似する上位 5 件の対話履歴とその共感応答を抽出する．類似度は，各対話履歴を Sentence-BERT[2]によりベクトル化し，それらベクトル間のコサイン類似度で算出する．その後，抽出した 5 個の対話事例をプロンプト（図 2 の {fewshot} 部分）に追加し，そのプロンプトを用いて gpt-4o で応答を生成する．

これら手法と知識ベースモデルの統合は，プロンプトを変更することで行う．図 2 は，知識ベースに 2 段階手法と Few-shot 手法を統合したモデルのプロンプトである．統合しない場合は該当する指示を除いたプロンプトを用いる．

3 実験

評価データは，JEmpatheticDialogue[1] のテストデータにおいて，対話履歴の最後が話し手の発話であるデータの中から無作為に抽出した 50 個のデータを使用した．評価は，Hu ら [3] のように，評価データに対して各モデルで生成した応答を，LLM により評価した．評価項目は，Qian ら [4] に倣い，共感性，一貫性，情報性，流暢性の 4 項目とし，1 から 5（高い値ほど良いことを示す）の 5 段階で評価した．各応答生成モデル及び評価で用いる LLM は，AzureOpenAIService のモデル ID:gpt-4o-2024-05-13 を使用した．temperature の値は 0 に設定した．

実験結果を表 1 に示す．なお，各値は，全評価データに対する評価結果の平均値である．ベースラインは，常識知識グラフによる推論，LLM によるユーザの感情や状況予測，Few-shot のいずれも行わずに gpt-4o で応答を生成させたモデルである．有意差検定はウィルコクソンの符号順位検定（有意水準 5%）で行った．表 1 中の「*」は，知識ベースとの差が統計的に有意であることを表している．

表 1 より，知識ベースが共感性の最高値を達成した．特に，知識ベースとベースラインの差は統計的に有意であった．このことから，常識知識グラフを活用することで日本語の共感応答生成の性能が向上することが確認できた．

また，知識ベースと 2 段階は，共感性と情報性において Few-shot よりも有意に優れていた．この理由としては，Few-shot では類似例を参考にするだけであるが，2 段階では 1 段階目で推定したユーザの感情や状況を，知識ベースでは関係「反応」の推論として感情や状況に類する情報を応答生成時に直接利用できたためと考えられる．一方で，ユーザの感情や状況に関する情報を直接利用する知識ベースと 2 段階は同等の性能であった．

「知識 +2 段階」は知識ベースと同等の性能であり，その他の統合モデルは知識ベースより共感性が低くなった．LLM はプロンプト中の情報を反映しようとするため，統合前の性能が低いノイズとなりうる情報を含めると，そのノイズに引っ張られてしまう可能性がある．したがって，単純に同一プロンプトに情報を追加する統合方法では性能改善につながらなかった可能性があると考えられる．

4 おわりに

本研究では，日本語において，常識知識グラフを活用した LLM による共感応答生成モデルを構築し，日本語の共感応答生成における常識知識グラフの有用性を検討した．JEmpatheticDialogue の 50 個のデータに対する応答性能を LLM により評価した結果，常識知識グラフを活用することで共感性が統計的に有意に改善することを確認した．一方で，知識ベースモデルは 2 段階手法のモデルと同等の性能であり，両者をプロンプトの中で組み合わせても応答の質は改善しなかった．今後は，人手による大規模な評価実験で更なる検証を行いたい．

参考文献

[1] H. Sugiyama et al. Empirical Analysis of Training Strategies of Transformer-based Japanese Chit-chat Systems. In *Proc. of IEEE SLT 2022*, pp. 685–691, 2023.

[2] S. Sabour et al. CEM: Commonsense-aware Empathetic Response Generation. In *Proc. of AAAI-22*, pp. 11229–11237, 2022.

[3] Y. Hu et al. APTNESS: Incorporating Appraisal Theory and Emotion Support Strategies for Empathetic Response Generation. In *Proc. of CIKM 2024*, pp. 900–909, 2024.

[4] Y. Qian et al. Harnessing the Power of Large Language Models for Empathetic Response Generation: Empirical Investigations and Improvements. In *Proc. of Findings of ACL: EMNLP 2023*, pp. 6516–6528, 2023.

[5] 井手他. 人間と言語モデルに対するプロンプトを用いたゼロからのイベント常識知識グラフ構築. 言語処理学会第 29 回年次大会発表論文集, pp. 322–327, 2023.

[2] https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2